p-Adic numbers in bioinformatics: from genetic code to PAM-matrix

A.Yu.Khrennikov[*], S.V.Kozyrev[†]

April 4, 2009

Abstract

In this paper we denonstrate that the use of the system of 2-adic numbers provides a new insight to some problems of genetics, in particular, generacy of the genetic code and the structure of the PAM matrix in bioinformatics. The 2-adic distance is an ultrametric and applications of ultrametrics in bioinformatics are not surprising. However, by using the 2-adic numbers we match ultrametric with a number theoretic structure. In this way we find new applications of an ultrametric which differ from known up to now in bioinformatics.

We obtain the following results. We show that the PAM matrix A allows the expansion into the sum of the two matrices $A = A^{(2)} + A^{(\infty)}$, where the matrix $A^{(2)}$ is 2–adically regular (i.e. matrix elements of this matrix are close to locally constant with respect to the discussed earlier by the authors 2–adic parametrization of the genetic code), and the matrix $A^{(\infty)}$ is sparse. We discuss the structure of the matrix $A^{(\infty)}$ in relation to the side chain properties of the corresponding amino acids.

1 Introduction

Various clustering procedures play a crucial role in bioinformatics, in particular, genetics, see, e.g., [1, 2] or [3]. An important class of such procedures is based on introduction of various metrics on the space information strings, see e.g. [5]. A metric with new interesting features was recently used in theoretical physics (from string theory to theory of disordered systems, spin glasses), see e.g. [6], [7], [8], [9] in cognitive science, psychology and image analysis [10]. This is so called p–adic metric (in fact, a class of metrics depending on the parameter p — a prime number). The main distinguishing feature of this metric is its sensitivity to hierarchic patterns in information having a special structure matching with p–adic encoding of information.[1]

A few years ago 2–adic metric was applied to study the problem of degeneration of the genetic code, see [11, 12, 13]. These p–adic models can be considered as new development in the approach to investigation of the structure of the genetic code from the point of view of coding theory, see [14, 15, 16].

In the present paper we discuss the structure of the PAM matrix used in bioinformatics (see for example [1]) from the point of view of p–adic analysis. We use the 2–adic parametrization of the genetic (amino acid) code obtained in [11] (see also [12] for the different p–adic parametrization).

In [11, 12] it was shown that, after some special parametrization of the space of codons (triples of nucleotides) the genetic code becomes a locally constant map of p–adic argument. Moreover, the degeneracy of the genetic code in this language takes the form of local constancy of the corresponding mapping.

Let us also mention the application of the p–adic parametrization to the description of the Parisi matrix from the replica symmetry breaking approach to spin glasses [17, 18]. After the p–adic parametrization of the numbers of the lines and the columns the Parisi matrix becomes a locally constant block matrix.

It is natural to check, using the p–adic parametrization approach, the structure of the PAM matrix. The PAM matrix is used in bioinformatics for sequence alignment and is constructed using a Markov chain model of point mutations for a protein chain.

[*] International Center for Mathematical Modelling in Physics and Cognitive Sciences, University of Växjö, S-35195, Sweden, e–mail: Andrei.Khrennikov@vxu.se

[†] Steklov Mathematical Institute, Moscow, Russia, e–mail: kozyrev@mi.ras.ru

[1] We remark that to appeal to hierarchical structures is quite common in genetics and bioinformatics in general, see, e.g. [3], [4]. Our main contribution is combining the hierarchic structure approach with number theory.

We assume that the structure of the PAM matrix has some relation to the structure of the genetic code. Using this idea we enumerate the lines and the columns of the PAM matrix using the 2–adic parametrization of the genetic code. After this parametrization the PAM matrix becomes more regular, namely, the dependence of the matrix elements A_{ij} of the PAM matrix on the indices i and j is close to locally constant with respect to the 2–adic norm for the majority of matrix elements.

We have some exceptions from this rule. It is easy to see that these exceptions are related to several amino acids, namely to Y, W, C, F, L. In order to describe this deviations from 2–adicity we introduce the following construction: we expand (by hands) the PAM matrix into the sum of the two matrices

$$A = A^{(2)} + A^{(\infty)}.$$

The matrix in this expansion $A^{(2)}$ is 2–adically regular (close to locally constant). The matrix $A^{(\infty)}$ is sparse (the majority of matrix elements are zero, non zero matrix elements are mainly concentrated of the lines and columns related to the amino acids Y, W, C, F, L).

One can see that the deviations from 2–adicity (i.e. non–zero matrix elements of $A^{(\infty)}$) are related to amino acids which are in some sense special — to the aromatic amino acids Y, W, F, and to Cysteine C which contains the SH group.

We also mention that the 2–adic structure of the genetic code is related to some chemical properties of the amino acids. In particular, hydrophobic amino acids are clustered in two ball with respect to the 2–adic norm. Therefore the 2–adic parametrization allows to separate the impact of the chemical and geometrical properties of aromatic amino acids for the structure of the PAM matrix.

The structure of the present paper is the following.

In Section 2 we discuss some family of ultrametric spaces.

In Section 3 we describe the 2–adic 2–dimensional parametrization of the genetic code of [11].

In Section 4 we put the PAM250 matrix.

In Section 5 we describe the reshuffling of the lines and the columns of the PAM matrix, corresponding to the 2–adic parametrization of the genetic code of Section 2.

In Section 6 we introduce the expansion of the PAM matrix into the sum of the two matrices, one of which is 2–adically regular (close to locally constant) and the other is sparse (majority of matrix elements are equal to zero).

Sections 7 and 8 are appendices where the definitions of PAM matrices and the eucaryotic genetic code are exposed.

2 Ultrametric spaces

An *ultrametric space* is a metric space where the metric $d(x,y)$ satisfies the strong triangle inequality:

$$d(x,y) \leq \max\left(d(x,z), d(y,z)\right), \qquad \forall x,y,z.$$

The strong triangle inequality can be stated geometrically: *each side of a triangle is at most as long as the longest one of the two other sides.* Such a triangle is quite restricted when considered in the ordinary Euclidean space — it is *isosceles,* i.e., $d(x,y) = d(y,z)$ or $d(x,z) = d(y,z)$ or $d(x,y) = d(z,x)$.

An ultrametric space is a natural mathematical object for description of a *hierarchical system.* On ultrametric spaces there exist many locally constant functions, i.e., functions which are constant on some vicinity of any point, but not necessarily constant on the whole space. In particular, we show that the genetic code can be considered as a locally constant map on a specially designed ultrametric space, so called 2-adic plane, see [11].

Let (X,d) be an ultrametric space. We consider balls $U_r(a) = \{x \in X : d(x,a) \leq r\}, r > 0, x \in X$. So, a is the center of the ball $U_r(a)$ having radius r. We mention a few unusual (from the viewpoint of usual Euclidean geometry) properties of ultrametric balls:

a). Each point of $U_r(a)$ can be chosen as its center. So, inside a ball all points have ”equal rights”.

b). Any two balls either do not intersect or one of the balls contains the other ball. In this framework, clustering into *disjoint balls* is a very natural operation.

We remark that ultrametric spaces were widely used in bioinformatics, see [1, 2]. However, in this paper we plan to elaborate new applications of ultrametric spaces to biology (genetics) which are different from mentioned ones.

We are interested in the following special class of ultrametric spaces (X, d). Every point x is the infinite sequence of digits

$$x = (\alpha_0, \alpha_1, \ldots, \alpha_n, \ldots) \,. \tag{1}$$

Each digit yields the finite number of values,

$$\alpha_i \in A_m = \{0, \ldots, m-1\}, \tag{2}$$

where $m > 1$ is a natural number, the base of the alphabet A_m.

If the sequence $x = (\alpha_0, \alpha_1, \ldots, \alpha_n, \ldots)$ contains only finite number of non-zero terms $(\alpha_0, \alpha_1, \ldots, \alpha_n)$, then we can consider x as the natural number

$$x = \sum_{i=0}^{n} \alpha_i m^i. \tag{3}$$

Moreover, this formula defines the one to one correspondence between natural numbers (with zero) and the space of final sequences $x = (\alpha_0, \alpha_1, \ldots, \alpha_n)$.

We denote the space of sequences (1), (2) by the symbol $\mathbb{Z}_m$. Ultrametric is introduced on this set in the following way. For two points

$$x = (\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n, \ldots), \qquad y = (\beta_0, \beta_1, \beta_2, \ldots, \beta_n, \ldots),$$

we set

$$d_m(x, y) = \frac{1}{m^k} \text{ if } \alpha_j = \beta_j, j = 0, 1, \ldots, k-1, \text{ and } \alpha_k \neq \beta_k.$$

The ultrametric space $(X = \mathbb{Z}_m, d = d_m)$ is called the space of m-adic integers.

The ultrametric d_m describes the following hierarchical structure. If $x = (\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots)$, $\alpha_j = 0, 1, \ldots, m-1$, is a vector encoding information on some object, then digits α_j have different weights. The digit α_0 is the most important, α_1 dominates over $\alpha_2, \ldots, \alpha_n, \ldots$, and so on. Such hierarchic information vectors can be created by living systems, e.g., by the brain to process information. Applications of m-adic numbers to information theory and, in particular, to description of cognitive processes and complex social systems were developed in [7, 10, 19, 20, 21, 22, 23].

In applications we will use not only "one dimensional" m-adic spaces, but also cartesian products of a few spaces, e.g., m-adic plane $\mathbb{Z}_m^2 = \mathbb{Z}_m \times \mathbb{Z}_m$ and so on. Our aim is to show that 2-adic plane structure was embedded in the genetic code, see [11].

We remark that m-adic numbers for $m = p$, where p is a prime number were intensively used (during last 20 years) in mathematical physics [6, 9]. The number theoretic definition is as follows.

Let us fix a prime number $p > 1$. For example, fix $p = 2$ or fix $p = 1999$. The example of ultrametric space is the field of p-adic numbers $\mathbb{Q}_p$, which is the completion of the field of rational numbers with respect to p-adic norm $|x|_p$, defined as follows: for a rational number $x = p^\gamma \frac{m}{n}$, where $\gamma = 0, \pm 1, \pm 2, \ldots$, and m, n are non-zero and are not divisible by p, its p-adic norm is

$$|x|_p = p^{-\gamma}.$$

p-Adic norm is widely used in number theory and algebraic geometry.

If a rational number is divisible by p^γ, where γ is very large, then its p-adic norm is very small. This hierarchy of the degrees of p (i.e. of values of the p–adic norm) gives the hierarchical (ultrametric) structure of the p–adic norm. p-Adic numbers are in one to one correspondence with the series

$$x = \sum_{i=\gamma}^{\infty} x_i p^i, \qquad x_i = 0, \ldots, p-1,$$

where γ is integer and $x_\gamma \neq 0$ (this expansion is the analogue of (3)).

The unit ball in $\mathbb{Q}_p$ with center at $a = 0$ zero coincides with the space of p-adic integers $\mathbb{Z}_p$.

p-Adic numbers were actively used (since pioneer papers of I.Volovich, [24]) in high energy physics (superstring theory, cosmology) and theory of disordered systems (spin glasses), see, e.g., review [9] and the book [6].

3 2-Adic parametrization of the genetic code

In the 2–adic parametrization approach of [11] we enumerate in some special way the set of 64 codons (triples of nucleotides) by pairs of digits (x, y), $x, y = 0, 1, 2, \dots, 7$. These digits are in one to one correspondence with the triples (x_0, x_1, x_2) and (y_0, y_1, y_2) of 0 and 1 (the expansions of x and y over degrees of 2):

$$x = x_0 + 2x_1 + 4x_2, \qquad y = y_0 + 2y_1 + 4y_2, \qquad x_i, y_i = 0, 1.$$

2–Adic norm of x is equal to 2^{-i}, where i is the number of the first non zero x_i in the above expansion (analogously for y).

Each pair of digits (x_i, y_i) is defined by a nucleotide using the rule

$$\begin{array}{|c|c|}\hline A & G \\ \hline U & C \\ \hline \end{array} = \begin{array}{|c|c|}\hline 00 & 01 \\ \hline 10 & 11 \\ \hline \end{array}.$$

Since the nucleotides $A = (0, 0)$ and $G = (0, 1)$ are purines, $U = (1, 0)$ and $C = (1, 1)$ are pyrimidines, the different first digits in the above binary representation corresponds to the different chemical types of the nucleotides. Namely, the nucleotide (x, y) with $x = 0$ is a purine, and the nucleotide (x, y) with $x = 1$ is a pyrimidine.

The second digit $y = 0, 1$ in the considered parametrization describes the H–bonding character (weak for $y = 0$ and strong for $y = 1$).

Using the above correspondence between the nucleotides and the digits 0 and 1, we introduce the correspondence between the codons (triples of nucleotides) and the triples (x_0, x_1, x_2) and (y_0, y_1, y_2) of 0 and 1 (equivalently, the corresponding $x, y = 0, 1, 2, \dots, 7$) by the following prescription.

The second nucleotide in the codon defines the pair (x_0, y_0), the first nucleotide in the codon defines the pair (x_1, y_1), and the third nucleotide in the codon defines the pair (x_2, y_2).

This rule is related to the following hierarchy of nucleotides in the codon

$$2 > 1 > 3$$

i.e the second nucleotide in the codon is the most important (and the largest in the 2–adic norm) and the third nucleotide is the least important.

After that we make the special reshuffling of the values of x and y:

$$0, 4, 2, 6, 1, 5, 3, 7 \mapsto 1, 2, 3, 4, 5, 6, 7, 8.$$

similar to made in [17].

Then we enumerate codons using the described above rule. Namely we put the codons in the table 8×8 with the natural 2–adic norm (here the numbers of the lines and columns are x and y defined above):

AAA	AAG	GAA	GAG	AGA	AGG	GGA	GGG
AAU	AAC	GAU	GAC	AGU	AGC	GGU	GGC
UAA	UAG	CAA	CAG	UGA	UGG	CGA	CGG
UAU	UAC	CAU	CAC	UGU	UGC	CGU	CGC
AUA	AUG	GUA	GUG	ACA	ACG	GCA	GCG
AUU	AUC	GUU	GUC	ACU	ACC	GCU	GCC
UUA	UUG	CUA	CUG	UCA	UCG	CCA	CCG
UUU	UUC	CUU	CUC	UCU	UCC	CCU	CCC

The 2×2 quadrates here are 2–dimensional 2–adic balls (or clusters) of the diameter 1/4.

After application of the eucaryotic genetic (amino acid) code (described in the Appendix) to the above table we get the table of amino acids on the 2–adic plane

$\frac{\text{K}}{\text{N}}$	$\frac{\text{E}}{\text{D}}$	$\frac{\text{R}}{\text{S}}$	G
$\frac{\text{Ter}}{\text{Y}}$	$\frac{\text{Q}}{\text{H}}$	$\frac{\text{Ter}\vert\text{W}}{\text{C}}$	R
$\frac{\text{I}\vert\text{M}}{\text{I}}$	V	T	A
$\frac{\text{L}}{\text{F}}$	L	S	P

where Ter is the stop codon. Each square in the above table is the 2 by 2 square in the 2-adic plane of codons. In particular, the genetic code map acts as follows on the 2–adic balls

$$\boxed{\begin{array}{cc} AAA & AAG \\ AAU & AAC \end{array}} \to \boxed{\begin{array}{c} \text{K} \\ \hline \text{N} \end{array}}, \qquad \boxed{\begin{array}{cc} CCA & CCG \\ CCU & CCC \end{array}} \to \boxed{\text{P}}$$

Here we use the standard notations for the nucleotides A, U, G, C and the amino acids.

We see, that the degeneracy of the genetic code in the above 2-adic parametrization is described by the 2-adic proximity — the codons which encode the same amino acid are 2-adically close. Moreover, the domains with the different degeneracy are symmetric at the 2-adic plane — the lower right half of the plane is occupied by amino acids with the degeneracy four, and the upper left half of the plane contains the amino acids with the degeneracy mainly equal to two. We also have the five cases of additional degeneracy which is not described by the 2-adic parametrization.

The described here 2-adic 2-dimensional parametrization of the genetic code is related to physical–chemical properties of the amino acids. Namely, the hydrophobic amino acids are clustered in the following two 2-adic balls in the 2-adic plane:

—	—	—	
—	—	$\frac{\text{Ter}\vert\text{W}}{\text{C}}$	
$\frac{\text{I}\vert\text{M}}{\text{I}}$	V		
$\frac{\text{L}}{\text{F}}$	L		

Here the hydrophobic amino acids are listed according to the book [25].

Using the 2-adic parametrization of the genetic code we can divide all the set of amino acid in the groups {K, N, E, D, Y, Q, H}, {R, G, W, C}, {I, M, V, L, F}, {T, A, S, P}. These groups are the images with respect to the map of the genetic code of the four quadrants (2–adic balls of the diameter 1/2) of the 2–adic plane of codons.

4 The PAM matrix

The following table describes the PAM250 matrix:

$$
A=\begin{array}{|c|cccccccccccccccccccc|}
\hline
\ast & \mathrm{A} & \mathrm{C} & \mathrm{D} & \mathrm{E} & \mathrm{F} & \mathrm{G} & \mathrm{H} & \mathrm{I} & \mathrm{K} & \mathrm{L} & \mathrm{M} & \mathrm{N} & \mathrm{P} & \mathrm{Q} & \mathrm{R} & \mathrm{S} & \mathrm{T} & \mathrm{V} & \mathrm{W} & \mathrm{Y} \\
\hline
\mathrm{A} & 2 & -2 & 0 & 0 & -4 & 1 & -1 & -1 & -1 & -2 & -1 & 0 & 1 & 0 & -2 & 1 & 1 & 0 & -6 & -3 \\
\mathrm{C} & \ast & 12 & -5 & -5 & -4 & -3 & -3 & -2 & -5 & -6 & -5 & -4 & -3 & -5 & -4 & 0 & -2 & -2 & -8 & 0 \\
\mathrm{D} & \ast & \ast & 4 & 3 & -6 & 1 & 1 & -2 & 0 & -4 & -3 & 2 & -1 & 2 & -1 & 0 & 0 & -2 & -7 & -4 \\
\mathrm{E} & \ast & \ast & \ast & 4 & -5 & 0 & 1 & -2 & 0 & -3 & -2 & 1 & -1 & 2 & -1 & 0 & 0 & -2 & -7 & -4 \\
\mathrm{F} & \ast & \ast & \ast & \ast & 9 & -5 & -2 & 1 & -5 & 2 & 0 & -4 & -5 & -5 & -4 & -3 & -3 & -1 & 0 & 7 \\
\mathrm{G} & \ast & \ast & \ast & \ast & \ast & 5 & -2 & -3 & -2 & -4 & -3 & 0 & -1 & -1 & -3 & 1 & 0 & -1 & -7 & -5 \\
\mathrm{H} & \ast & \ast & \ast & \ast & \ast & \ast & 6 & -2 & 0 & -2 & -2 & 2 & 0 & 3 & 2 & -1 & -1 & -2 & -3 & 0 \\
\mathrm{I} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 5 & -2 & 2 & 2 & -2 & -2 & -2 & -2 & -1 & 0 & 4 & -5 & -1 \\
\mathrm{K} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 5 & -3 & 0 & 1 & -1 & 1 & 3 & 0 & 0 & -2 & -3 & -4 \\
\mathrm{L} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 6 & 4 & -3 & -3 & -2 & -3 & -3 & -2 & 2 & -2 & -1 \\
\mathrm{M} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 6 & -2 & -2 & -1 & 0 & -2 & -1 & 2 & -4 & -2 \\
\mathrm{N} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 2 & -1 & 1 & 0 & 1 & 0 & -2 & -4 & -2 \\
\mathrm{P} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 6 & 0 & 0 & 1 & 0 & -1 & -6 & -5 \\
\mathrm{Q} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 4 & 1 & -1 & -1 & -2 & -5 & -4 \\
\mathrm{R} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 6 & 0 & -1 & -2 & 2 & -4 \\
\mathrm{S} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 2 & 1 & -1 & -2 & -3 \\
\mathrm{T} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 3 & 0 & -5 & -3 \\
\mathrm{V} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 4 & -6 & -2 \\
\mathrm{W} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 17 & 0 \\
\mathrm{Y} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 10 \\
\hline
\end{array}
$$

Because the matrix A is symmetrical we put only the half of matrix elements. This matrix looks irregular and has no any obvious structure. In the next Section we will show that some reshuffling of the numbers of the lines and columns will put this matrix in more regular form.

5 Reshuffling of matrix elements of the PAM matrix

Let us enumerate the lines and the columns of the PAM matrix A of the previous Section using the 2–adic parametrization of the genetic code. We get for A the following:

$$
A=\begin{array}{|c|ccccccc|cccc|ccccc|cccc|}
\hline
\ast & \mathrm{K} & \mathrm{N} & \mathrm{E} & \mathrm{D} & \mathrm{Y} & \mathrm{Q} & \mathrm{H} & \mathrm{R} & \mathrm{G} & \mathrm{W} & \mathrm{C} & \mathrm{I} & \mathrm{M} & \mathrm{V} & \mathrm{L} & \mathrm{F} & \mathrm{T} & \mathrm{A} & \mathrm{S} & \mathrm{P} \\
\hline
\mathrm{K} & 5 & 1 & 0 & 0 & -4 & 1 & 0 & 3 & -2 & -3 & -5 & -2 & 0 & -2 & -3 & -5 & 0 & -1 & 0 & -1 \\
\mathrm{N} & 1 & 2 & 1 & 2 & -2 & 1 & 2 & 0 & 0 & -4 & -4 & -2 & -2 & -2 & -3 & -4 & 0 & 0 & 1 & -1 \\
\mathrm{E} & 0 & 1 & 4 & 3 & -4 & 2 & 1 & -1 & 0 & -7 & -5 & -2 & -2 & -2 & -3 & -5 & 0 & 0 & 0 & -1 \\
\mathrm{D} & 0 & 2 & 3 & 4 & -4 & 2 & 1 & -1 & 1 & -7 & -5 & -2 & -3 & -2 & -4 & -6 & 0 & 0 & 0 & -1 \\
\mathrm{Y} & -4 & -2 & -4 & -4 & 10 & -4 & 0 & -4 & -5 & 0 & 0 & -1 & -2 & -2 & -1 & 7 & -3 & -3 & -3 & -5 \\
\mathrm{Q} & 1 & 1 & 2 & 2 & -4 & 4 & 3 & 1 & -1 & -5 & -5 & -2 & -1 & -2 & -2 & -5 & -1 & 0 & -1 & 0 \\
\mathrm{H} & 0 & 2 & 1 & 1 & 0 & 3 & 6 & 2 & -2 & -3 & -3 & -2 & -2 & -2 & -2 & -2 & -1 & -1 & -1 & 0 \\
\hline
\mathrm{R} & 3 & 0 & -1 & -1 & -4 & 1 & 2 & 6 & -3 & 2 & -4 & -2 & 0 & -2 & -3 & -4 & -1 & -2 & 0 & 0 \\
\mathrm{G} & -2 & 0 & 0 & 1 & -5 & -1 & -2 & -3 & 5 & -7 & -3 & -3 & -3 & -1 & -4 & -5 & 0 & 1 & 1 & -1 \\
\mathrm{W} & -3 & -4 & -7 & -7 & 0 & -5 & -3 & 2 & -7 & 17 & -8 & -5 & -4 & -6 & -2 & 0 & -5 & -6 & -2 & -6 \\
\mathrm{C} & -5 & -4 & -5 & -5 & 0 & -5 & -3 & -4 & -3 & -8 & 12 & -2 & -5 & -2 & -6 & -4 & -2 & -2 & 0 & -3 \\
\hline
\mathrm{I} & -2 & -2 & -2 & -2 & -1 & -2 & -2 & -2 & -3 & -5 & -2 & 5 & 2 & 4 & 2 & 1 & 0 & -1 & -1 & -2 \\
\mathrm{M} & 0 & -2 & -2 & -3 & -2 & -1 & -2 & 0 & -3 & -4 & -5 & 2 & 6 & 2 & 4 & 0 & -1 & -1 & -2 & -2 \\
\mathrm{V} & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -1 & -6 & -2 & 4 & 2 & 4 & 2 & -1 & 0 & 0 & -1 & -1 \\
\mathrm{L} & -3 & -3 & -3 & -4 & -1 & -2 & -2 & -3 & -4 & -2 & -6 & 2 & 4 & 2 & 6 & 2 & -2 & -2 & -3 & -3 \\
\mathrm{F} & -5 & -4 & -5 & -6 & 7 & -5 & -2 & -4 & -5 & 0 & -4 & 1 & 0 & -1 & 2 & 9 & -3 & -4 & -3 & -5 \\
\hline
\mathrm{T} & 0 & 0 & 0 & 0 & -3 & -1 & -1 & -1 & 0 & -5 & -2 & 0 & -1 & 0 & -2 & -3 & 3 & 1 & 1 & 0 \\
\mathrm{A} & -1 & 0 & 0 & 0 & -3 & 0 & -1 & -2 & 1 & -6 & -2 & -1 & -1 & 0 & -2 & -4 & 1 & 2 & 1 & 1 \\
\mathrm{S} & 0 & 1 & 0 & 0 & -3 & -1 & -1 & 0 & 1 & -2 & 0 & -1 & -2 & -1 & -3 & -3 & 1 & 1 & 2 & 1 \\
\mathrm{P} & -1 & -1 & -1 & -1 & -5 & 0 & 0 & 0 & -1 & -6 & -3 & -2 & -2 & -1 & -3 & -5 & 0 & 1 & 1 & 6 \\
\hline
\end{array}
$$

Here we divided all the set of amino acids to the groups {K, N, E, D, Y, Q, H}, {R, G, W, C}, {I, M, V, L, F}, {T, A, S, P}. These four groups correspond to the second nucleotide in the codons which encode (through the amino acid genetic code) amino acids in the corresponding group. Namely, the first group correspond to the codons with A (Adenine) at the second position, the second group correspond to the codons with G (Guanine), the third group correspond to the codons with U (Uracil), and the fourth group correspond to the codons with C (Cytosine).

Compared to the PAM matrix from the previous Section, this reshuffled matrix is more regular. It has large positive matrix elements at the main diagonal, and off diagonal terms are close to block constant at least at some of the 16 blocks, corresponding to matrix elements with the indices from one of the 4 groups of amino acids.

In particular, in the block corresponding to the matrix elements A_{ij}, $i, j \in$ {T, A, S, P}, all off diagonal matrix elements are equal to 1 (10 matrix elements) or 0 (2 matrix elements).

In the block A_{ij}, $i \in$ {K, N, E, D, Y, Q, H}, $j \in$ {T, A, S, P} we have matrix elements equal to 0 (13 matrix elements), -1 (10 matrix elements), 1 (1 matrix element) and anomalous matrix elements equal to -3 and -5 corresponding to the amino acid Y. The analogous situation we will have in the other blocks of the matrix A.

We arrive to the following picture: the PAM matrix A will be a block matrix with matrix elements close to locally constant (i.e. constant in the 16 blocks) if we will exclude matrix elements corresponding to some amino acids, namely, to the amino acids Y, W, C, L, F, and R.

6 Expansion for the PAM matrix

In the present Section we introduce the main construction of this paper: we will expand the PAM matrix A in the sum of the matrices $A^{(2)}$ and $A^{(\infty)}$:

$$A = A^{(2)} + A^{(\infty)},$$

where the matrix $A^{(2)}$ will be 2–adically regular (matrix elements are close to locally constant), and the matrix $A^{(\infty)}$ will be sparse (i.e. majority of matrix elements of this matrix will be equal to zero).

We propose the following choice for matrices $A^{(2)}$ and $A^{(\infty)}$.

$$A^{(2)} = \begin{array}{|c|ccccccc|cccc|ccccc|cccc|}
\hline
* & K & N & E & D & Y & Q & H & R & G & W & C & I & M & V & L & F & T & A & S & P \\
\hline
K & 5 & 1 & 1 & 1 & 1 & 1 & 0 & -1 & -2 & -1 & -1 & -2 & -2 & -2 & -2 & -2 & 0 & -1 & 0 & -1 \\
N & 1 & 2 & 1 & 2 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & -2 & -2 & 0 & 0 & 0 & -1 \\
E & 1 & 1 & 4 & 3 & 1 & 2 & 1 & -1 & 0 & -1 & -1 & -2 & -2 & -2 & -2 & -2 & 0 & 0 & 0 & -1 \\
D & 1 & 2 & 3 & 4 & 1 & 2 & 1 & -1 & 0 & -1 & -1 & -2 & -3 & -2 & -2 & -2 & 0 & 0 & 0 & -1 \\
Y & 1 & 1 & 1 & 1 & 10 & 1 & 0 & -1 & -1 & 0 & 0 & -1 & -2 & -2 & -1 & -2 & -1 & -1 & -1 & -1 \\
Q & 1 & 1 & 2 & 2 & 1 & 4 & 3 & -1 & -1 & -1 & -1 & -2 & -1 & -2 & -2 & -2 & -1 & 0 & -1 & 0 \\
H & 0 & 2 & 1 & 1 & 0 & 3 & 6 & -1 & -2 & -1 & -1 & -2 & -2 & -2 & -2 & -2 & -1 & -1 & -1 & 0 \\
\hline
R & -1 & 0 & -1 & -1 & -1 & -1 & -1 & 6 & -3 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -1 & -2 & 0 & 0 \\
G & -2 & 0 & 0 & 0 & -1 & -1 & -2 & -3 & 5 & -3 & -3 & -3 & -3 & -1 & -2 & -3 & 0 & -1 & 1 & -1 \\
W & -1 & 0 & -1 & -1 & 0 & -1 & -1 & -2 & -3 & 17 & -3 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\
C & -1 & 0 & -1 & -1 & 0 & -1 & -1 & -2 & -3 & -3 & 12 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -3 \\
\hline
I & -2 & -2 & -2 & -2 & -1 & -2 & -2 & -2 & -3 & -2 & -2 & 5 & 2 & 2 & 2 & 2 & 0 & -1 & -1 & -2 \\
M & -2 & -2 & -2 & -3 & -2 & -1 & -2 & -2 & -3 & -2 & -2 & 2 & 6 & 2 & 2 & 2 & -1 & -1 & -2 & -2 \\
V & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -1 & -2 & -2 & 2 & 2 & 4 & 2 & 2 & 0 & 0 & -1 & -1 \\
L & -2 & -2 & -2 & -2 & -1 & -2 & -2 & -2 & -3 & -2 & -2 & 2 & 2 & 2 & 6 & 2 & -1 & -1 & -2 & -2 \\
F & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -3 & -2 & -2 & 2 & 2 & 2 & 2 & 9 & -1 & -1 & -2 & -2 \\
\hline
T & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & -2 & -2 & 0 & -1 & 0 & -1 & -1 & 3 & 1 & 1 & 0 \\
A & -1 & 0 & 0 & 0 & -1 & 0 & -1 & -2 & -1 & -2 & -2 & -1 & -1 & 0 & -1 & -1 & 1 & 2 & 1 & 1 \\
S & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 1 & -2 & -2 & -1 & -2 & -1 & -2 & -2 & 1 & 1 & 2 & 1 \\
P & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -2 & -3 & -2 & -2 & -1 & -2 & -2 & 0 & 1 & 1 & 6 \\
\hline
\end{array}$$

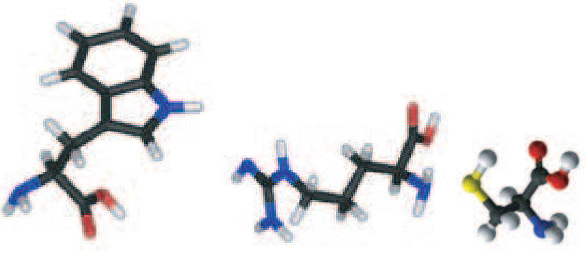

Figure 1: tryptophan W, arginine R, cysteine C

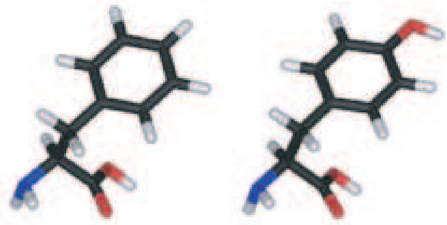

Figure 2: phenylalanine F, tyrosine Y

$$A^{(\infty)} = \begin{array}{|c|ccccccc|cccc|ccccc|cccc|}
\hline
* & K & N & E & D & Y & Q & H & R & G & W & C & I & M & V & L & F & T & A & S & P \\
\hline
K & 0 & 0 & -1 & -1 & -5 & 0 & 0 & 4 & 0 & -2 & -4 & 0 & 2 & 0 & -1 & -3 & 0 & 0 & 0 & 0 \\
N & 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 & -4 & -4 & 0 & 0 & 0 & -1 & -2 & 0 & 0 & 1 & 0 \\
E & -1 & 0 & 0 & 0 & -5 & 0 & 0 & 0 & 0 & -6 & -4 & 0 & 0 & 0 & -1 & -3 & 0 & 0 & 0 & 0 \\
D & -1 & 0 & 0 & 0 & -5 & 0 & 0 & 0 & 1 & -6 & -4 & 0 & 0 & 0 & -2 & -4 & 0 & 0 & 0 & 0 \\
Y & -5 & -3 & -5 & -5 & 0 & -5 & 0 & -3 & -4 & 0 & 0 & 0 & 0 & 0 & 0 & 9 & -2 & -2 & -2 & -4 \\
Q & 0 & 0 & 0 & 0 & -5 & 0 & 0 & 2 & 0 & -4 & -4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
R & 4 & 0 & 0 & 0 & -3 & 2 & 3 & 0 & 0 & 4 & -2 & 0 & 2 & 0 & -1 & -2 & 0 & 0 & 0 & 0 \\
G & 0 & 0 & 0 & 1 & -4 & 0 & 0 & 0 & 0 & -4 & 0 & 0 & 0 & 0 & -1 & -2 & 0 & 2 & 0 & 0 \\
W & -2 & -4 & -6 & -6 & 0 & -4 & -2 & 4 & -4 & 0 & -5 & -3 & -2 & -4 & 0 & 2 & -3 & -4 & 0 & -4 \\
C & -4 & -4 & -4 & -4 & 0 & -4 & -2 & -2 & 0 & -5 & 0 & 0 & -3 & 0 & -4 & -2 & 0 & 0 & 2 & 0 \\
\hline
I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 2 & 0 & -1 & 0 & 0 & 0 & 0 \\
M & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & -2 & -3 & 0 & 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\
V & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4 & 0 & 2 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \\
L & -1 & -1 & -1 & -2 & 0 & 0 & 0 & -1 & -1 & 0 & -4 & 0 & 2 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\
F & -3 & -2 & -3 & -4 & 9 & -3 & 0 & -2 & -2 & 2 & -2 & -1 & -2 & -3 & 0 & 0 & -2 & -3 & -1 & -3 \\
\hline
T & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 & -1 & -2 & 0 & 0 & 0 & 0 \\
A & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 2 & -4 & 0 & 0 & 0 & 0 & -1 & -3 & 0 & 0 & 0 & 0 \\
S & 0 & 1 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\
P & 0 & 0 & 0 & 0 & -4 & 0 & 0 & 0 & 0 & -4 & 0 & 0 & 0 & 0 & -1 & -3 & 0 & 0 & 0 & 0 \\
\hline
\end{array}$$

Non zero matrix elements of $A^{(\infty)}$ are mainly concentrated on the lines and columns corresponding to Y, W, C, L, F. There are also several non–zero matrix elements corresponding to R and some other amino acids.

We see that the non zero matrix elements of $A^{(\infty)}$ are mainly concentrated on aromatic amino acids, such as Y, F, W, and on C which contains the SH group. Therefore deviations from 2–adic regularity (i.e. the block structure of the $A^{(2)}$ matrix) can be discussed as related to the geometric properties of the side chains of amino acids (for aromatic amino acids Y, F, W, and for Arginine R), and to the ability of Cysteine C to create a disulfide bond.

The above pictures show the amino acids W, R, C, F, Y corresponding to non-zero matrix elements of the matrix $A^{(\infty)}$. Let us mention that amino acids F and Y for which the corresponding matrix element $A^{(\infty)}_{FY}$ is very large (and majority of the other matrix elements of $A^{(\infty)}$ for these amino acids are negative) are very similar from the point of view of geometry.

Of course, our analysis of PAM matrix based on the 2-adic plane representation of the genetic code is only the first step in using p-adic numbers in genetics and bioinformatics in general. We hope to proceed towards other important problems, cf., e.g., [1]– [4], [26]. Finally, we mention the famous "In Silico Biology" project, see, e.g., Yamato et al. [27]. We point out that, in fact, operation of any computer can be represented as 2-adic dynamical system, see [10], [28]. Therefore 2-adic representation of the genetic code might be useful in realization of the "In Silico Biology" project.

7 Appendix: the PAM matrix

In this section we discuss the construction of the Dayhoff PAM matrix, which can be found for example in [1, 2]. We start with blocks — ungapped multiple alignments of proteins from existing databases. Any sequence in the block is no more than 15% different from any other sequence in this block.

Then, the Markov model, which reproduces the mentioned blocks of proteins was constructed. This Markov model is defined by the amino acid substitutions (point mutations). We have the stationary distribution p_a of the probabilities of the amino acids, $\sum_{a=1}^{20} p_a = 1$, and the transition probability p_{ab}, normalized by the condition that the probability of a point mutation (substitution of the amino acid) at one step of the Markov model is equal to 0.01:

$$\sum_{a,b=1}^{20} p_{ab} p_b = 0.01.$$

Then we take the matrix given by the n steps of the Markov model, i.e. the n-th degree P^n of the matrix $P = (p_{ab})$, and consider the matrix with the matrix elements

$$A^{(n)} = \log_{10}\left(\frac{(P^n)_{ab}}{p_b}\right).$$

This matrix is known as the PAM matrix (usually n is taken to be equal to 250 and the matrix elements are approximated by integers).

8 Appendix: the genetic code

The following table describes the eucaryotic genetic code — the correspondence between codons (triples of nucleotides and amino acids):

AAA	K	UAA	Ter	GAA	E	CAA	Q
AAU	N	UAU	Y	GAU	D	CAU	H
AAG	K	UAG	Ter	GAG	E	CAG	Q
AAC	N	UAC	Y	GAC	D	CAC	H
AUA	I	UUA	L	GUA	V	CUA	L
AUU	I	UUU	F	GUU	V	CUU	L
AUG	M	UUG	L	GUG	V	CUG	L
AUC	I	UUC	F	GUC	V	CUC	L
AGA	R	UGA	Ter	GGA	G	CGA	R
AGU	S	UGU	C	GGU	G	CGU	R
AGG	R	UGG	W	GGG	G	CGG	R
AGC	S	UGC	C	GGC	G	CGC	R
ACA	T	UCA	S	GCA	A	CCA	P
ACU	T	UCU	S	GCU	A	CCU	P
ACG	T	UCG	S	GCG	A	CCG	P
ACC	T	UCC	S	GCC	A	CCC	P

Acknowledgments The authors would like to thank B.Dragovich, I.V.Volovich for fruitful discussions and valuable comments. The authors were partially supported by the grant of International center for mathematical modeling in physics, engineering and cognitive science, University of Vaxjo. One of the authors (A.K.) was partially supported by the QPIC grant, Tokyo University of Science. One of the authors (S.K.) gratefully acknowledges being partially supported by the grants DFG Project 436 RUS 113/809/0-1 and DFG Project 436 RUS 113/951, by the grants of The Russian Foundation for Basic Research RFFI 05-01-04002-NNIO-a and RFFI 08-01-00727-a, by the grant of the President of Russian Federation for the support of scientific schools NSh-3224.2008.1, by the Program of the Department of Mathematics of Russian Academy of Science "Modern problems of theoretical mathematics" and by the program of Ministry of Education and Science of Russia "Development of the scientific potential of High School, years of 2009–2010", project 3341.

References

[1] R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.

[2] A.Isaev, *Introduction to mathematical methods in bioinformatics*, Springer, 2006.

[3] M. Nishihama, Yu. Sakatsuji, A. Arinami, and S. Miyazaki, Informational approach for the study of cis-regulatory elements and DNA binding proteins. In: L. Accardi, W. Freudenberg, M. Ohya (eds.), Quantum Bio-Informatics, pp. 371– 380. WSP, Singapore, 2007.

[4] T. Suzuki and S. Miyazaki, Basics of genome sequence analysis in bioinformatics - its fundamental ideas and problems In: L. Accardi, W. Freudenberg, M. Ohya (eds.), Quantum Bio-Informatics, pp. 299 – 313. WSP, Singapore, 2008.

[5] F.Murtagh, A.Heck, *Multivariate Data Analysis*, Kluwer Academic Publishers, Dordrecht, 1987.

[6] V.S. Vladimirov, I.V. Volovich, Ye.I. Zelenov, *p–Adic analysis and mathematical physics*, World Scientific, Singapore, 1994 (See also Nauka, Moscow, 1994, in Russian).

[7] A. Khrennikov, *Non–Archimedean Analysis: Quantum Paradoxes, Dynamical Systems and Biological Models*, Kluwer Academic Publishers, 1997.

[8] S.V. Kozyrev, *Methods and applications of ultrametric and p–adic analysis: from wavelet theory to biophysics*. Modern problems of mathematics. Issue 12. Steklov Mathematical Institute, Moscow, 2008, (in Russian) http://www.mi.ras.ru/spm/pdf/012.pdf.

[9] B. Dragovich, A. Yu. Khrennikov, S. V. Kozyrev and I. V. Volovich, On p-adic mathematical physics. p-Adic Numbers, Ultrametric Analysis and Applications, **1**, N 1, 1-17 (2009).

[10] A.Yu. Khrennikov, *Information dynamics in cognitive, psychological and anomalous phenomena*, Series in Fundamental Theories of Physics, Kluwer, Dordrecht, 2004.

[11] A.Yu. Khrennikov, S.V. Kozyrev, Genetic code on the diadic plane // Physica A: Statistical Mechanics and its Applications. 2007. V.381. P.265-272. arXiv:q-bio.QM/0701007

[12] B.Dragovich, A.Dragovich, A p-Adic Model of DNA Sequence and Genetic Code, p-Adic Numbers, Ultrametric Analysis and Applications, **1**, N 1, 34-41 (2009). arXiv:q-bio/0607018v1

[13] A.Yu. Khrennikov, p–Adic information space and gene expression. In: Integrative approaches to brain complexity, eds. S.Grant, N.Heintz, J.Noebels, Welcome Truct Publ. P.14. 2006.

[14] R.Swanson, A unifying concept for the amino acid code, Bulletin of Mathematical Biology, 1984. V.46. N.2. P.187-203.

[15] M.Sjöstrom, S.Wold, A multivariate study of the relationship between the genetic code and the physical–chemical properties of amino acids. Journal of Molecular Evolution. 1985. V.22. P.272-277.

[16] M.D.Perlwitz, C.Burks, M.S.Waterman, Pattern Analysis of the Genetic Code, Advances in applied mathematics, 1988. V.9. P.7-21.

[17] V.A.Avetisov, A.H.Bikulov, S.V.Kozyrev, Application of p–adic analysis to models of spontaneous breaking of replica symmetry, // J. Phys. A: Math. Gen. 1999. V.32. N.50. P.8785–8791, arXiv:cond-mat/9904360

[18] G.Parisi, N.Sourlas, p–Adic numbers and replica symmetry breaking // European Phys. J. B. 2000. V.14. P.535–542. arXiv:cond-mat/9906095

[19] A. Yu. Khrennikov, Probabilistic pathway representation of cognitive information. *J. Theor. Biology,* **231**, 597-613 (2004).

[20] A. Yu. Khrennikov, p-adic discrete dynamical systems and collective behaviour of information states in cognitive models. *Discrete Dynamics in Nature and Society,* **5,** 59-69 (2000).

[21] S.Albeverio, A.Yu.Khrennikov, P.Kloeden, Memory retrieval as a p-adic dynamical system. Biosystems, 49, 105-115 (1999).

[22] D.Dubischar, V.M.Gundlach, O.Steinkamp, A. Yu.Khrennikov, A p-adic model for the process of thinking disturbed by physiological and information noise. J. Theor. Biology,197, 451-467 (1999).

[23] A.Yu. Khrennikov, Human subconscious as the p-adic dynamical system. J. of Theor. Biology. 193, 179-196 (1998).

[24] I.V.Volovich, p-Adic string, Class. Quantum Gravity. 1987. V.4. L.83-L87.
I.V.Volovich, Number theory as the ultimate physical theory. Preprint No. TH 4781/87, CERN, Geneva, 1987.

[25] A.V.Finkelshtein, O.B.Ptitsyn, *Physics of Proteins*, Academic Press, London, 2002.

[26] D. Wanke, J. Killan, A basic introduction to gene expression studies using microarray expression data analysis. In: L. Accardi, W. Freudenberg, M. Ohya (eds.), Quantum Bio-Informatics, pp. 314 – 326. WSP, Singapore, 2008.

[27] I. Yamato, T. Ando, A. Suzuki, K. Harada, S. Itoh, S. Miyazaki, N. Kobayashi, M. Takeda, Toward In Silico Biology (from sequences to systems). In: L. Accardi, W. Freudenberg, M. Ohya (eds.), Quantum Bio-Informatics, pp. 440 – 455. WSP, Singapore, 2007.

[28] V. Anashin and A. Yu. Khrennikov, *Applied algebraic dynamics.* De Gruyter, Berlin (2009).